Titre du projet : ***Évolution d'une application ESG – Projet Streamlit ESG***

Nom : FAVEUR
Prénom : Kenson

Université : Paris Cité

Formation : Analyste Data Science

Résumé

Ce projet s'inscrit dans la dynamique croissante d'intégration des critères ESG (Environnement, Social, Gouvernance) dans l'analyse de la performance des entreprises. À travers le développement d'une application interactive sous Streamlit, enrichie par un modèle de réseau de neurones, l'objectif a été de rendre les scores ESG accessibles, lisibles et prédictibles. Après une phase d'exploration et de traitement des données issues de la base S&P 500 ESG Risk Ratings, un moteur d'analyse prédictive a été mis en œuvre, permettant d'anticiper le score ESG global d'une entreprise sur la base de ses caractéristiques sectorielles et financières. L'application permet ainsi à l'utilisateur d'explorer les performances ESG, de comparer plusieurs entreprises, et de simuler des scénarios futurs. Les résultats obtenus, avec un R^2 de 0.9825 et une erreur moyenne faible, confirment la pertinence du modèle développé.

Introduction

Face à la montée des préoccupations sociétales et environnementales, les critères ESG (Environnement, Social, Gouvernance) se sont imposés comme des indicateurs majeurs dans l'évaluation des entreprises. Institutions financières, analystes, investisseurs, ONG et même citoyens s'appuient de plus en plus sur ces données pour orienter leurs décisions. Pourtant, ces indicateurs restent souvent complexes à interpréter, difficilement accessibles ou réservés à des experts.

Dans ce contexte, la science des données et l'intelligence artificielle deviennent des leviers puissants pour démocratiser l'accès aux informations ESG. Le présent projet s'inscrit dans cette dynamique : il vise à concevoir une application interactive de visualisation et d'analyse des scores ESG d'entreprises à travers une interface développée avec Streamlit, renforcée par l'intégration d'un modèle de réseau de neurones.

L'objectif initial était de bâtir un tableau de bord simple, capable de présenter de manière visuelle les performances ESG d'entreprises. Rapidement, le projet a évolué vers une approche plus intelligente : la prédiction des scores ESG à partir de données d'entrée grâce à l'apprentissage automatique. En utilisant un réseau de neurones entraîné sur des données historiques et financières, combiné à des métriques statistiques robustes , l'application propose non seulement une lecture des performances passées mais aussi une anticipation potentielle des évolutions ESG.

L'enjeu est double :

Pédagogique : permettre aux étudiants, analystes ou curieux d'approfondir leur compréhension de l'ESG via des outils modernes.

Ce rapport présente donc le cheminement du projet : de l'idée initiale à l'intégration d'un moteur d'analyse prédictive et conversationnelle, en passant par la conception d'une interface Streamlit dédiée. Il détaillera également les choix technologiques, les étapes de mise en œuvre, les limites rencontrées ainsi que les perspectives offertes par cette application ESG augmentée par l'IA.

Contexte de l'organisation ou du cadre du projet

Ce projet est réalisé dans le cadre de la formation Analyste Data Science, dispensée à l'université Paris Cité. Il s'inscrit plus précisément dans un module orienté data science et intelligence artificielle appliquée, visant à initier les étudiants aux problématiques réelles du traitement et de l'analyse de données dans un contexte professionnel ou sociétal. Ce module encourage les apprenants à concevoir des solutions technologiques autour de problématiques actuelles, avec un accent particulier sur la responsabilité numérique et l'impact social de la donnée.

L'objectif pédagogique est double : d'une part, permettre aux étudiants de maîtriser un outil de développement d'applications interactives (ici, Streamlit), et d'autre part, les initier aux techniques de modélisation prédictive via l'apprentissage automatique. Le projet sur les données ESG a été retenu en raison de sa pertinence actuelle et de son potentiel de transversalité entre technologie, finance, développement durable et gouvernance d'entreprise.

Le choix du thème ESG (Environnement, Social, Gouvernance) répond à plusieurs enjeux fondamentaux. Premièrement, la durabilité est au cœur des préoccupations modernes. Les entreprises, qu'elles soient cotées ou non, sont de plus en plus évaluées sur leur capacité à intégrer les principes du développement durable dans leurs opérations. Deuxièmement, ces données sont souvent complexes à interpréter pour un public non expert. Les rendre visuellement accessibles et intelligibles devient alors un levier éducatif et stratégique. Troisièmement, l'analyse ESG constitue un terrain d'application idéal pour tester des algorithmes prédictifs et des approches d'intelligence artificielle, car elle repose sur des données multivariées, dynamiques et souvent hétérogènes.

Enfin, ce projet s'inscrit dans une perspective pédagogique innovante qui dépasse la simple production d'un outil technique : il vise à faire réfléchir les étudiants sur les usages concrets de l'IA dans la société, sur la manière dont on traite et restitue des indicateurs sensibles, et sur le rôle de la visualisation dans la démocratisation de l'information complexe.

Présentation du projet existant

L'application actuellement développée avec Streamlit vise à rendre accessibles les données ESG (Environnement, Social, Gouvernance) à un public élargi via une interface web interactive. Ce projet combine la puissance de la visualisation de données, le traitement statistique avancé, et l'intelligence artificielle pour permettre à l'utilisateur de mieux comprendre les performances non financières des entreprises.

Fonctionnalités principales de l'application

L'application Streamlit est articulée autour des fonctions suivantes :

Sélection dynamique d'entreprises : L'utilisateur peut choisir une ou plusieurs entreprises depuis une liste déroulante alimentée automatiquement par une base de données. Cette sélection déclenche le traitement des données correspondantes.

Affichage de scores ESG individuels : Une fois une entreprise sélectionnée, ses scores Environnemental, Social et Gouvernance sont affichés sous forme de barres horizontales pour une lecture simple. Une note ESG globale est également présentée.

Comparaison entre entreprises :

Une section dédiée permet de comparer les scores de deux entreprises sélectionnées côte à côte. Les résultats sont affichés en double graphique, facilitant l'analyse comparative rapide.

Un module IA est prévu pour générer une analyse textuelle synthétique de la performance ESG de chaque entreprise, basée sur les données et les prévisions du modèle de réseau de neurones.

Données exploitées

La base de données utilisée pour alimenter cette application provient du site Kaggle. Elle est accessible à l'adresse suivante : S&P 500 ESG Risk Ratings Dataset. Ce jeu de données fournit des évaluations des risques ESG (Environnement, Social, Gouvernance) pour les entreprises composant l'indice S&P 500. Il contient des indicateurs tels que les scores environnementaux, sociaux et de gouvernance, permettant d'analyser la performance durable des entreprises cotées.

Structure du jeu de données

Ce jeu de données contient les indicateurs suivants :

- **Symbol** : Code boursier (ticker) de l'entreprise.
- **Full Name** : Nom complet de la société cotée au S&P 500.
- **GICS Sector** : Secteur d'activité selon la classification GICS (Global Industry Classification Standard), ex. “Information Technology”.
- **GICS Sub-Industry** : Sous-secteur GICS précisant davantage le domaine d'activité.
- **environmentScore** : Score relatif aux pratiques et impacts environnementaux (émissions, ressources, etc.).
- **socialScore** : Score mesurant la performance sociale (droits humains, diversité, relation client, etc.).
- **governanceScore** : Score de gouvernance (structure du conseil, éthique, transparence, etc.).
- **totalEsg** : Score ESG global combinant les trois dimensions précédentes.
- **highestControversy** : Niveau maximal de controverse associé à l'entreprise (échelle d'exposition au risque médiatique ou éthique).
- **percentile** : Positionnement relatif de l'entreprise par rapport aux autres, sous forme de centile ESG.
- **ratingYear** et **ratingMonth** : Date d’attribution du score ESG.
- **marketCap** : Capitalisation boursière de l’entreprise (en milliards ou millions selon le format).
- **beta** : Coefficient de volatilité par rapport au marché (indicateur financier de risque).
- **overallRisk** : Score agrégé de risque ESG (combinaison des performances et des controverses).

Prétraitement des données ESG

Avant toute exploration ou modélisation, les données ont été soumises à un processus rigoureux de prétraitement afin d’assurer leur qualité, leur cohérence et leur pertinence analytique. Ce traitement s’est déroulé en plusieurs étapes :

Nettoyage et préparation

- Vérification de l’absence de valeurs manquantes sur l’ensemble des variables sélectionnées.
- Standardisation des noms de colonnes et des types de données.
- Filtrage éventuel des valeurs aberrantes (par exemple, capitalisation boursière excessivement élevée ou valeur nulle de score ESG total).

- Conversion des champs temporels (ratingYear, ratingMonth) en un format exploitable pour une éventuelle analyse chronologique.

Sélection des variables pertinentes

Un sous-ensemble de variables numériques a été conservé pour l'étude des corrélations :

- environmentScore
- socialScore
- governanceScore
- totalEsg
- highestControversy
- percentile
- marketCap
- beta
- overallRisk

Ces indicateurs couvrent les dimensions environnementales, sociales et de gouvernance, ainsi que des informations de risque, de performance relative et de données financières.

Analyse de corrélation

Une matrice de corrélation a été générée afin d'évaluer les relations linéaires entre la variable cible totalEsg et les autres indicateurs. L'objectif était d'identifier :

- Les dimensions ESG les plus contributives au score global,
- L'impact potentiel des controverses (highestControversy et overallRisk),
- Le lien éventuel entre performance extra-financière et capitalisation boursière (marketCap, beta).

Nouveau concept issu des cours : Réseaux de neurones et métriques de performance

Dans le cadre de la formation suivie, un module a été consacré à la **modélisation prédictive via les réseaux de neurones artificiels**, un concept issu du domaine du machine learning, particulièrement adapté aux problématiques complexes de régression et de classification.

Au cours de ce module, plusieurs notions clés ont été explorées :

- **Structure du réseau de neurones** L'architecture d'un réseau repose sur des **couches d'entrée, des couches cachées** (hidden layers) et une **couche de sortie**. Le choix du **nombre de neurones** dans les couches cachées influence directement la capacité du modèle à détecter des motifs complexes dans les données. Une structure trop simple

peut conduire à un sous-apprentissage (*underfitting*), tandis qu'un modèle trop complexe peut surapprendre (*overfitting*).

- **Hyperparamètres d'entraînement**
    - **Nombre d'epochs** : il s'agit du nombre de fois que l'intégralité des données d'entraînement est présentée au modèle. Ce paramètre permet de contrôler la durée et la profondeur de l'apprentissage.
    - **Taux de dropout** : mécanisme de régularisation qui consiste à désactiver aléatoirement un pourcentage de neurones pendant l'apprentissage, dans le but de renforcer la robustesse du modèle. Un taux typique peut varier entre 0.2 et 0.5 selon la complexité du jeu de données.
- **Évaluation du modèle : métriques de régression** Plusieurs indicateurs ont été utilisés pour évaluer les performances du modèle lors de la prédiction de variables ESG :
    - **MAE (Mean Absolute Error)** : moyenne des écarts absolus entre les valeurs réelles et prédites.
    - **MSE (Mean Squared Error)** : moyenne des carrés des écarts. Plus sensible aux grandes erreurs.
    - **R^2 (coefficient de détermination)** : mesure statistique indiquant la proportion de variance expliquée par le modèle. Une valeur proche de 1 indique une bonne précision du modèle.

Cette composante pédagogique a permis non seulement de mieux comprendre la logique d'apprentissage automatique, mais aussi de l'appliquer concrètement à travers une application Streamlit intégrant un réseau de neurones destiné à prédire un score ESG global à partir de variables sectorielles et financières.

Évolution proposée du projet à partir des nouveaux apports

L'enrichissement du projet repose sur l'intégration de concepts avancés issus de la formation, en particulier autour des réseaux de neurones, de l'interprétabilité des résultats, et de la valorisation interactive des données ESG à travers Streamlit. Les évolutions proposées visent à renforcer la pertinence analytique, l'autonomie utilisateur et la capacité prédictive de l'application :

Ajout d'un moteur prédictif basé sur un réseau de neurones

Un réseau de neurones a été intégré pour modéliser la variable totalEsg, en exploitant des données environnementales, sociales, de gouvernance ainsi que des indicateurs financiers (ex. marketCap, beta, etc.). Ce modèle permet :

d'**estimer un score ESG probable** pour de nouvelles entreprises ou des scénarios fictifs ;

de simuler les impacts d'une variation de certains facteurs (ex. gouvernance, controverses) sur la performance ESG.

Évaluation et interprétation dynamique des résultats

L'interface Streamlit a été étendue pour afficher des **métriques clés** (MAE, MSE, R^2) permettant d'évaluer la qualité du modèle. En complément, des **résumés visuels et interprétables** sont proposés à l'aide de graphiques dynamiques (loss curve, comparaisons entre secteurs, etc.).

Interaction personnalisée avec les hyperparamètres

Les utilisateurs peuvent désormais ajuster :

- le **nombre d'epochs** d'entraînement,
- le **taux de dropout** utilisé pour la régularisation,
- et potentiellement le **nombre de neurones** dans la couche cachée.

Cela permet de transformer l'application en **espace expérimental pédagogique**, où chacun peut observer l'impact de ses choix sur la performance du réseau.

Enrichissement par l'extraction et l'interprétation de rapports PDF

Une fonctionnalité locale a été ajoutée pour **analyser automatiquement des rapports ESG au format PDF**, extraire les informations importantes, et permettre un dialogue utilisateur sous forme de chatbot intelligent basé sur ces documents.

Perspectives futures

Des extensions ambitieuses pourraient inclure :

- la création d'un **système de recommandation ESG** par secteur d'activité ;
- l'intégration de l'interprétabilité du modèle via des méthodes comme **SHAP** pour expliquer les prédictions ;
- l'extension à d'autres jeux de données ou secteurs pour une **généralisation multi-industries**.

Principales étapes du projet

Objectif général : concevoir une application interactive capable d'explorer, d'analyser et de prédire des scores ESG à partir de données structurées (Kaggle) et non structurées (rapports PDF), tout en intégrant un modèle de réseau de neurones pour la prédiction.

1. Cadrage du projet

- Définition des besoins fonctionnels et techniques.
- Identification des sources de données (Kaggle, rapports PDF, etc.).
- Choix des outils : Python, Streamlit, Keras, Pandas, etc.

2. Collecte et exploration des données

- Téléchargement du jeu de données depuis Kaggle.
- Inventaire et chargement des rapports ESG PDF.
- Analyse exploratoire : structure, types de données, distributions.

3. Prétraitement des données

- Nettoyage : gestion des valeurs manquantes, normalisation, transformation des types.
- Sélection des variables pertinentes
- Étude des corrélations entre les variables.

4. Modélisation prédictive (réseau de neurones)

- Définition du modèle : architecture, neurones, fonctions d'activation.
- Séparation du jeu de données (entraînement/test).
- Entraînement du modèle avec réglage des hyperparamètres (epochs, dropout).
- Évaluation des performances (MAE, MSE, R^2).

5. Développement de l'application Streamlit

- Création de l'interface utilisateur pour visualiser, explorer et prédire les scores ESG.
- Intégration du modèle prédictif.
- Intégration d'un module d'analyse automatique des rapports PDF.
- Affichage des métriques de performance et visualisations interactives.

6. Expérimentations et ajustements

- Tests utilisateurs, recueil de feedbacks.
- Amélioration continue : raffinage du modèle, ajustement de l'interface, ajout de filtres ESG avancés.

Analyse des coûts et gains engendrés par l'évolution du projet

L'évolution du projet vers une plateforme interactive intégrant un modèle de réseau de neurones, une interface Streamlit avancée et un module d'analyse documentaire entraîne des investissements techniques et humains. Toutefois, ces coûts sont largement compensés par les bénéfices attendus, tant sur le plan pédagogique que fonctionnel.

Coûts estimés

Poste de coût	Détail
Temps de développement	Implémentation du modèle, interface Streamlit, intégration PDF
Ressources techniques	Ordinateur performant, bibliothèques Python, stockage local
Courbe d'apprentissage	Temps nécessaire pour maîtriser les frameworks ML et Streamlit
Maintenance et amélioration continue	Ajustements du modèle, corrections de bugs, mises à jour de l'interface

Gains attendus

Type de gain	Détail
Pédagogique	Application concrète des notions de deep learning, NLP, visualisation
Fonctionnel	Prédiction ESG, interrogation de rapports PDF, visualisation dynamique
Explicabilité	Meilleure compréhension des facteurs ESG grâce à l'interprétation du modèle
Réutilisabilité	Possibilité d'adapter l'outil à d'autres jeux de données ou secteurs

Analyse et interprétation de la base de données ESG

Présentation du jeu de données

Le jeu de données provient de Kaggle : *S&P 500 ESG Risk Ratings*. Il recense les indicateurs ESG (Environnement, Social, Gouvernance) de plusieurs grandes entreprises cotées, avec des variables telles que :

- environmentScore, socialScore, governanceScore : dimensions ESG
- totalEsg : score global ESG

- marketCap, beta, overallRisk, etc.

Analyse descriptive des variables ESG et financières

L'analyse porte sur 426 observations représentant des entreprises cotées au S&P 500, évaluées selon différents critères ESG et financiers. Voici les enseignements clés issus des statistiques descriptives :

Scores ESG (environmentScore, socialScore, governanceScore, totalEsg)

Le score environnemental varie de 0 à 24.98 avec une moyenne relativement faible (≈5.78), indiquant que de nombreuses entreprises sont encore peu engagées sur cet axe.

Le score social est globalement plus élevé (moyenne ≈9.07), traduisant une prise en compte plus développée des enjeux liés aux droits humains, à la diversité ou à la qualité de vie au travail.

Le score de gouvernance est relativement stable (moyenne ≈6.70 ; écart-type modéré), témoignant d'une attention généralement constante portée à la structure de direction et à la transparence.

Le score ESG global (totalEsg) a une moyenne de 21.55 avec un maximum à 41.66, illustrant une large dispersion des performances durables entre les entreprises.

Niveau de controverse (highestControversy)

La moyenne s'élève à 1.88 (sur 5), ce qui suggère un risque modéré de réputation ou d'incidents éthiques. Toutefois, la présence de valeurs allant jusqu'à 5 montre que certaines entreprises sont exposées à des polémiques majeures.

Données financières (marketCap, beta)

La capitalisation boursière est très variable (de 6.4 milliards à 3.296 billions USD), avec une forte dispersion (écart-type très élevé), indiquant la cohabitation de géants et d'acteurs plus modestes.

Le coefficient bêta, indicateur de volatilité, est centré autour de 1.03. On observe des cas extrêmes (jusqu'à 3.24) signalant des entreprises très sensibles aux mouvements du marché.

Autres indicateurs

Le percentile ESG moyen est de 35.8, ce qui peut indiquer une concentration des valeurs vers le bas du classement (à confirmer par histogramme).

La variable overallRisk, notée de 1 (faible risque) à 10 (très élevé), a une moyenne de 5.29 : l'échantillon présente donc des profils ESG très hétérogènes.

Analyse de la corrélation avec la variable cible totalEsg

L'étude de la corrélation linéaire entre le score ESG global (totalEsg) et les autres variables du jeu de données révèle plusieurs enseignements importants :

Corrélations fortes et positives :

percentile (r = 0.99) Une quasi-redondance statistique avec totalEsg, ce qui est logique puisqu'il s'agit probablement d'un classement basé sur ce score. Cette variable pourrait même être retirée dans une analyse prédictive pour éviter la multicolinéarité.

environmentScore (r = 0.70) Forte corrélation : les entreprises performantes globalement en ESG ont très souvent un score environnemental élevé. Cela reflète le poids significatif des enjeux climatiques et écologiques dans l'évaluation globale.

socialScore (r = 0.70) Relation presque aussi forte que celle avec l'environnement. Cela montre que la responsabilité sociale (diversité, conditions de travail, droits humains) contribue largement au score ESG final.

Corrélations modérées :

highestControversy (r ≈ 0.40) Corrélation inversement intuitive : certaines entreprises très exposées aux controverses peuvent maintenir un score ESG correct, probablement grâce à des pratiques de compensation ou de communication efficaces. Cette variable capte bien la "face publique" du risque réputationnel.

governanceScore (r ≈ 0.35) Corrélation plus faible que pour l'environnement ou le social. Cela peut refléter une plus grande homogénéité des pratiques de gouvernance (ex. : obligations réglementaires similaires).

Corrélations faibles ou nulles :

overallRisk (r ≈ 0.09) Faible lien direct avec le score ESG, ce qui suggère que ce risque agrégé capture une autre dimension, peut-être perçue à travers les controverses ou les impacts financiers.

beta (r ≈ 0.03) et marketCap (r ≈ -0.01) Absence de relation significative : la volatilité boursière et la taille de l'entreprise n'ont pas d'effet direct sur la notation ESG globale.

Interprétation des métriques du modèle de prédiction ESG

MAE (Mean Absolute Error) = 0.6486

- Cela signifie que l'erreur moyenne entre les prédictions et les valeurs réelles du score ESG est **d'environ 0.65 point**.
- En d'autres termes, le modèle est en moyenne à moins de **1 point d'erreur** sur une échelle dont la valeur cible (totalEsg) peut aller jusqu'à 40+.
- C'est un **niveau d'erreur très faible**, indiquant une bonne précision.

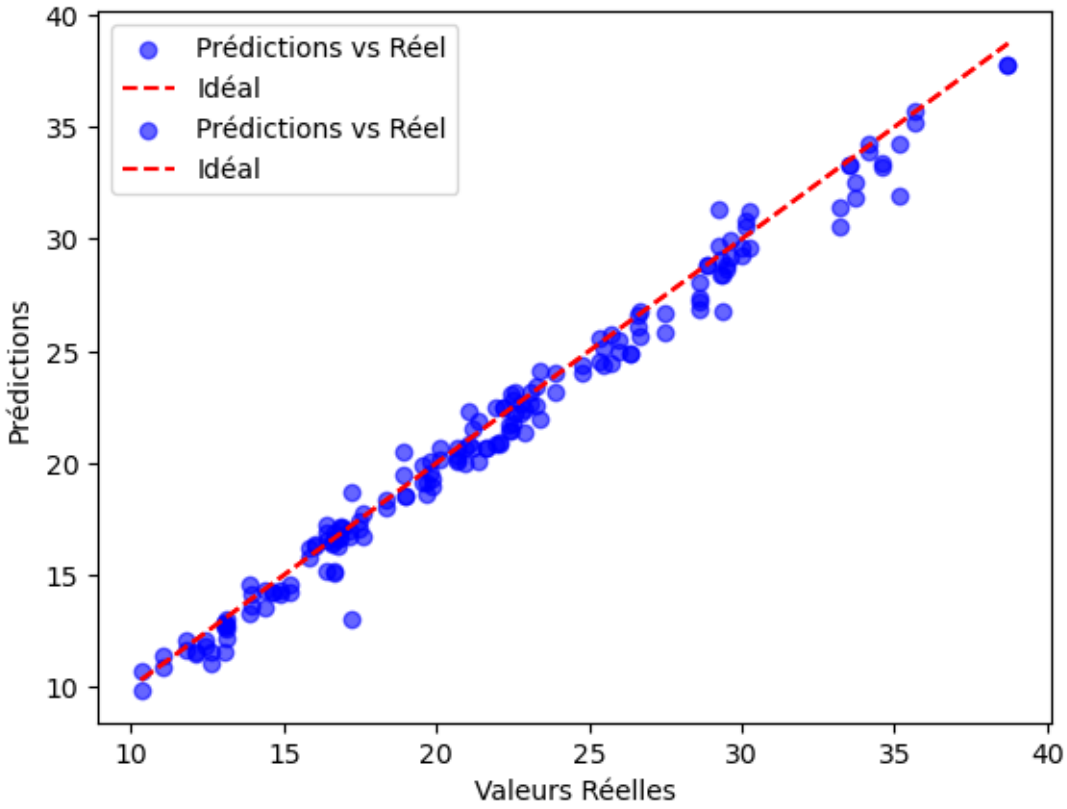
MSE (Mean Squared Error) = 0.7865

- L'erreur quadratique moyenne étant inférieure à 1, cela montre que **les erreurs importantes (valeurs extrêmes)** sont également maîtrisées.
- Comme le MSE est plus sensible aux grandes erreurs que le MAE, cette valeur confirme **l'homogénéité des performances** du modèle sur différents cas.

R^2 Score = 0.9825

- Le score de détermination R^2 mesure la proportion de la variance expliquée par le modèle.
- Une valeur de **0.98** signifie que le modèle explique **plus de 98 % des variations** du score ESG.
- Autrement dit, le modèle **généralise très bien** et capture les dépendances structurelles du jeu de données avec grande précision.

Graphe des résultats



Interprétation du graphique « Prédictions vs Valeurs Réelles »

Le graphique ci-dessus illustre la comparaison entre les prédictions du modèle et les valeurs réelles observées. Chaque point bleu représente une paire (valeur réelle, prédiction) générée par le modèle.

La ligne rouge pointillée correspond à la ligne idéale y=xy = xy=x, représentant une prédiction parfaitement exacte. L'alignement serré des points bleus autour de cette ligne indique que le modèle parvient à estimer les valeurs avec une grande précision.

On observe peu de dispersion autour de la ligne idéale, ce qui suggère une faible erreur de prédiction et une bonne généralisation du modèle. De plus, aucune tendance systématique de sur- ou sous-estimation n'est visible, ce qui indique l'absence de biais significatif.

Ce résultat met en évidence la capacité du modèle à fournir des prédictions fiables, tant pour les petites que pour les grandes valeurs de la variable cible.

Conclusion

Ce projet a démontré la capacité des technologies de data science et d'intelligence artificielle à renforcer la lisibilité et l'utilité des indicateurs ESG dans un contexte pédagogique et analytique. L'intégration d'un modèle de réseau de neurones dans une interface conviviale comme Streamlit permet de combiner rigueur technique et accessibilité utilisateur. L'application finale offre une double valeur : d'une part, elle constitue un outil d'aide à la décision en matière de durabilité ; d'autre part, elle représente une plateforme pédagogique dynamique pour expérimenter avec la modélisation prédictive et l'analyse ESG.

Les résultats très satisfaisants obtenus en termes de performance du modèle ($R^2 > 0.98$) confirment la robustesse de l'approche, tout en ouvrant la voie à des perspectives d'amélioration : intégration de méthodes d'interprétabilité (ex. SHAP), extension à d'autres indices boursiers, ou ajout de sources de données textuelles pour enrichir l'analyse qualitative.

Bibliographie

1. Bengio, Y., Courville, A., & Goodfellow, I. (2016). *Deep Learning*. MIT Press.

2. Chollet, F. (2017). *Deep Learning with Python*. Manning Publications.

3. ESG Risk Ratings Dataset, Kaggle. Disponible sur : https://www.kaggle.com/datasets

4. Streamlit Documentation. Disponible sur : https://docs.streamlit.io

5. Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python*, *Journal of Machine Learning Research*, 12, pp. 2825–2830.

6. UN PRI (Principles for Responsible Investment). https://www.unpri.org

7. Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press.

8. Kienner, J.-P. (2024). *Notes de cours – Réseaux de neurones et Deep Learning*. Université Paris Cité, Formation Analyste Data Science.